Near-end listening enhancement using a noise-robust linear time-invariant filter

Villani, Filippo; Chan, Wai-Yip; Tan, Zheng-Hua; Østergaard, Jan; Jensen, Jesper

Published in:
2024 18th International Workshop on Acoustic Signal Enhancement, IWAENC 2024 - Proceedings

DOI (link to publication from Publisher):
10.1109/IWAENC61483.2024.10694258

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

NEAR-END LISTENING ENHANCEMENT USING A NOISE-ROBUST LINEAR TIME-INVARIANT FILTER

Filippo Villani[1]*, Wai-Yip Chan*[2]*, Zheng-Hua Tan*[1]*, Jan Østergaard*[1]*, Jesper Jensen*[1,3]

[1]Aalborg University, Department of Electronic Systems, Aalborg, 9220, Denmark
[2]Queen's University, Department of Electrical and Computer Engineering, Kingston, K7L 3N6, Canada
[3]Oticon A/S, Smørum, 2765, Denmark

ABSTRACT

In environments with competing sound sources, speech intelligibility can be significantly compromised. This paper addresses the near-end listening enhancement (NELE) problem, i.e., the problem of processing an available clean speech signal in order to maximize its intelligibility when it is subsequently presented to a human listener in an adverse acoustic situation. We propose a time-invariant and low-complexity NELE algorithm that maximizes an approximation of the Speech Intelligibility Index by redistributing speech energy across frequency bands. Unlike existing algorithms, the proposed algorithm incorporates a mechanism that allows it to distinguish between temporally fluctuating and non-fluctuating noise maskers by using only long-term speech and noise statistics. Simulation results show that the proposed method outperforms baseline algorithms, whether time-invariant or time-varying, in a wide range of noise conditions.

Index Terms— Near-end listening enhancement, speech intelligibility, linear time-invariant filters

1. INTRODUCTION

Speech communication plays a fundamental role in every aspect of life. Unfortunately, it is often hindered by noise and reverberation, which may affect the speech intelligibility of the received signal. In this paper, we consider the situation where the clean signal is available for processing before playback in a noisy background, e.g., public address systems. This problem is commonly referred to as near-end listening enhancement (NELE) [1–4].

Many solutions to this problem find inspiration in how humans adapt their speech production to the environment [3] in order to effectively convey their message, a phenomenon known as the Lombard effect [5]. This speaking behavior occurs whenever humans speak in a noisy environment, and is achieved by increasing the vocal effort to produce an over-articulated speech signal that results in higher intelligibility compared to neutrally-produced speech presented in the same noise and SNR level [6]. Some of these algorithms apply time-invariant filters to the entire speech signal, with the general effect of reducing its spectral tilt by boosting higher frequencies, while maintaining the original speech energy [1, 7, 8]. More complex algorithms employ filters that can enhance the spectral contour of the fundamental frequency [9] or the formants [10]. Other methods divide the speech signal into consecutive time segments and apply different time-invariant filters for each segment, while maintaining the speech energy within each segment [11]. A large class of methods shift energy not only across frequency bands but also through time. This can be achieved, e.g., using dynamic range compression [10, 12–15], which increases intelligibility by boosting low-energy components, such as transient sounds, at the expense of higher energy components, like vowels. Some of these methods further improve intelligibility by modifying the duration of the uttered speech in an attempt to replicate this aspect of the Lombard effect [12, 13] or by minimizing the temporal overlap between the target speaker and a temporally fluctuating noise masker [16], i.e., a type of noise whose amplitude varies significantly over time. More recent work on NELE has focused on machine learning techniques such as Gaussian mixture models [17–19] and deep neural networks [20, 21] to learn the mapping of speech features from conversational to Lombard speech, or generative models [22] to learn and produce the desired speech modification end-to-end.

In this paper, we focus on low-complexity solutions, and propose a time-invariant and low-complexity method inspired by optimalASII [7]. As [7], the proposed method relies on an estimate of the per-critical-band SNR, but, unlike [7], this SNR is formulated as a function of long-term temporal statistics of the noise. Specifically, we modify the definition of the SNR by replacing the conventional noise power with a fractile noise power. Despite its simplicity, in simulation experiments, the proposed method outperforms baseline methods such as SEO [9], the best performing algorithm in the Hurricane Challenge 1 (HC-1) [2], and optimalASII [7].

2. REVIEW OF OPTIMALASII

In this section, we briefly review the original optimalASII algorithm [7], as the proposed algorithm is closely inspired by it. We note that optimalASII was named optimalSII in [23], but we use optimalASII to underline the fact that the objective function used is an approximation of the SII (ASII).

OptimalASII processes the clean speech signal through a linear time-invariant filter. Let $X(j,l)$ and $V(j,l)$ denote the results of passing a realization of the clean speech signal and the noise through an auditory filterbank, whose filters are gammatone filters with center frequencies equally spaced on the ERB scale and bandwidths of approximately 1 ERB. The indices j and l denote a sub-band- and a time-index, respectively. OptimalASII assumes access to the average energy of speech and noise within each sub-band, here named long-term power, $\sigma^2_{X_j} = 1/L_X \sum_{l=1}^{L_X} |X(j,l)|^2$ and $\sigma^2_{V_j} = 1/L_V \sum_{l=1}^{L_V} |V(j,l)|^2$, $j = 1, ..., J$, where L_X and L_V denote the number of speech and noise frames, respectively, and J is the number of sub-bands. These are then used to compute the long-term SNR ξ_j in each sub-band, defined as

$$\xi_j = \frac{\sigma^2_{X_j}}{\sigma^2_{V_j}}, \qquad j = 1, ..., J, \tag{1}$$

This work is partly funded by the William Demant Foundation.

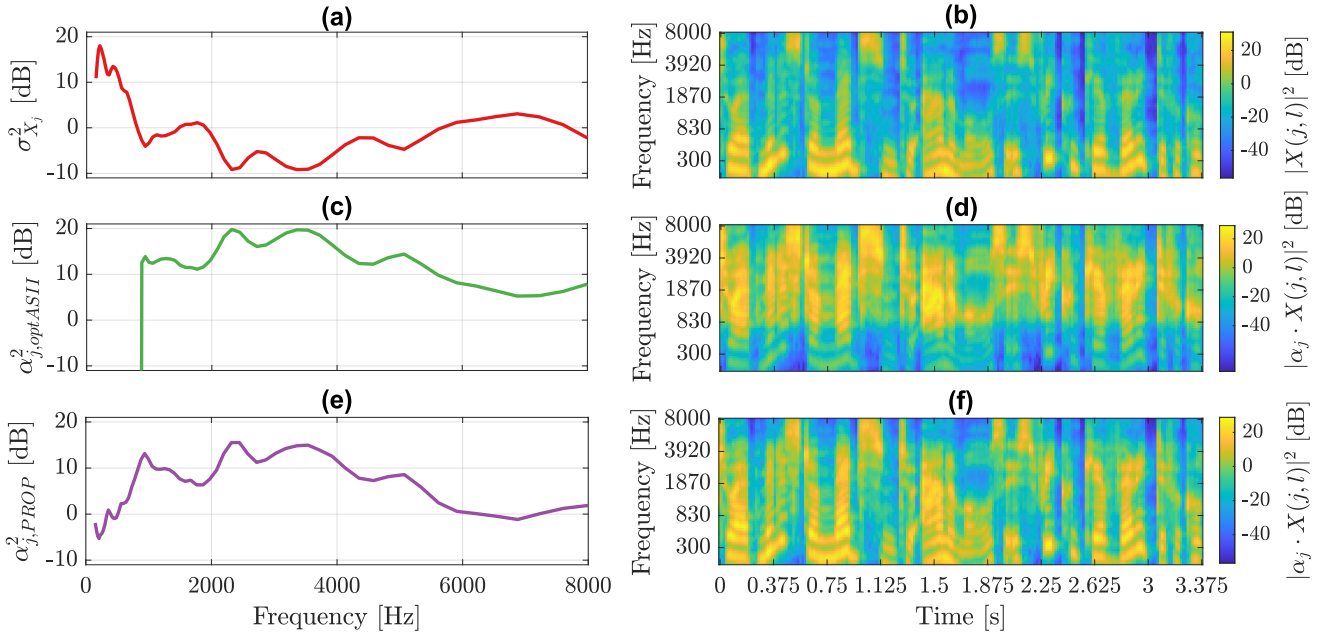

Fig. 1. (a) Speech long-term power and (b) its spectrogram. (c, e) sub-band gains of the optimalASII and the proposed methods, respectively, for SNR = -10 dB. (d, f) spectrograms of speech processed by the optimalASII and the proposed methods, respectively.

which are mapped to the ASII through the following relations [7]:

$$d(\xi_j) = \frac{\xi_j}{\xi_j + 1}, \tag{2}$$

$$\text{ASII} = \sum_{j=1}^{J} \gamma_j \cdot d(\xi_j), \tag{3}$$

where the scalar $0 \leq \gamma_j \leq 1$ is the value of the band-importance function in the j-th sub-band as defined by the SII standard [24]. OptimalASII then maximizes the estimated intelligibility in terms of ASII (as defined in (3)) subject to an energy constraint on the processed signal. More specifically, optimalASII finds the optimal set of sub-band gains $\alpha_j^{*^2}$ to apply to the clean speech:

$$\begin{aligned} \alpha_j^{*^2} &= \arg\max_{\alpha_j^2 \geq 0} \sum_{j=1}^{J} \gamma_j d(\alpha_j^2 \xi_j) \\ \text{s.t.} \quad & \sum_{j=1}^{J} \alpha_j^2 \sigma_{X_j}^2 = \sum_{j=1}^{J} \sigma_{X_j}^2, \\ & \alpha_j^2 \sigma_{X_j}^2 \geq 0, \forall j. \end{aligned} \tag{4}$$

Solving the problem leads to the following solution [7]:

$$\alpha_j^{*2} \sigma_{X_j}^2 = \max\left(\frac{\sigma_{V_j} \sqrt{\gamma_j}}{\sqrt{\nu}} - \sigma_{V_j}^2, 0 \right), \forall j, \tag{5}$$

where $\nu > 0$ is chosen to satisfy the energy constraint defined in (4):

$$\frac{1}{\sqrt{\nu}} = \frac{\sum_{j=1}^{J} \sigma_{X_j}^2 + \sum_{j \in \mathcal{M}} \sigma_{V_j}^2}{\sum_{j \in \mathcal{M}} \sqrt{\gamma}_j \sigma_{V_j}}, \tag{6}$$

and where $\mathcal{M}$ is the set of sub-bands whose optimal gains are positive. The solution of (4) can be implemented using a simple bisection method to find the appropriate value of $\sqrt{\nu}$ [7].

3. MOTIVATION AND FORMULATION OF PROPOSED ALGORITHM

In this section, we first provide a qualitative analysis of optimalASII in order to identify its potential weaknesses (Sec. 3.1). This serves as a motivation for the proposed method (Sec. 3.2).

3.1. Qualitative analysis of optimalASII

In HC-1 [2], optimalASII performed well in speech-shaped noise (SSN), showing comparable results to SEO [9], the best performing algorithm in the challenge. On the other hand, optimalASII performed significantly worse when a competing speaker (CS) was presented along with the target speech. In fact, in this scenario, listening tests revealed that optimalASII degraded speech intelligibility [2].

This degradation can be due to two reasons: first, it is well-known [25] that SII is inherently incapable of dealing with fluctuating noise, since its prediction depends only on long-term averages of the speech and noise energy, and thus temporal aspects of the noise are ignored. Second, initial informal listening tests of optimalASII-processed speech in the presence of CS and an analysis of the shape of its sub-band gains over frequency revealed that optimalASII tends to over-suppress frequency regions, especially in frequency bands below 800 Hz, where the local SNR is low. To illustrate this, Fig. 1 shows the long-term power $\sigma_{X_j}^2$ of speech along with its spectrogram (Fig. 1(a, b)), as well as the frequency response of optimalASII and the spectrogram of the processed speech signal (Fig. 1(c, d)) when speech and CS noise are mixed at an SNR of -10 dB. The over-suppression of low-frequency regions, which can be clearly observed in Fig. 1(c, d), is undesirable as they contain significant information about the target speech, e.g., the fundamental frequency and first formants [26] (cf. Fig. 1(b, d)), which can be used by the listener for speaker identification and speech understanding [27]. While in SSN these regions are severely degraded over the entire duration of the

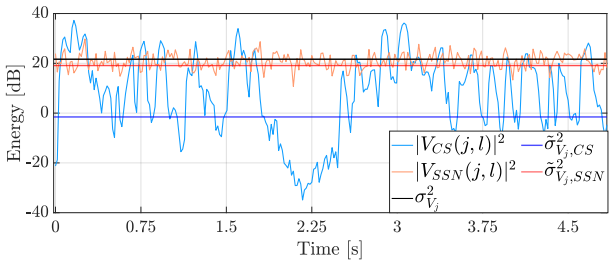

Fig. 2. Noise energy for SSN and CS in band $j = 15$ ($c_f = 522$ Hz) as a function of time along with the corresponding long-term fractile noise power $\tilde{\sigma}^2_{V_j}$ ($\phi = 0.3$) and the conventional long-term power $\sigma^2_{V_j}$. Note that while $\tilde{\sigma}^2_{V_j}$ is different for SSN and CS, the long-term noise power $\sigma^2_{V_j}$ is the same for both signals. For visualization purposes, the magnitudes are displayed on a dB scale.

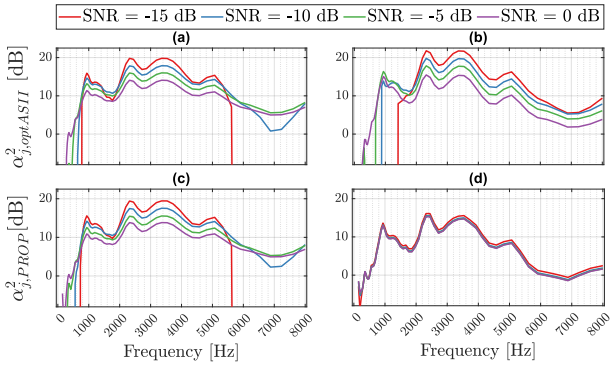

Fig. 3. Gains for each sub-band in the filterbank for different SNRs. (a) optimalASII in SSN and (b) CS, (c) proposed method ($\phi = 0.3$) in SSN and (d) CS.

target speech signal, in fluctuating noise, such as CS, the level of the noise masker may be low enough to momentarily unmask the target speech [28]. This is a well-known phenomenon, often referred to as "listening in the dips" or "glimpsing" [16, 29].

3.2. Proposed method

Motivated by this analysis, we pursue the idea of developing an algorithm which behaves like optimalASII in SSN, but avoids the over-suppression in fluctuating noise (cf. Fig. 1(e)) to allow for glimpsing.

An immediate solution to dealing with fluctuating noise would be to allow the model to process the clean speech signal in a time-varying manner, e.g., where the estimate of the sub-band SNR ξ_j and the optimal gain values α^2_j are computed for and applied to successive time segments [11]. However, time-frequency processing often leads to unpleasant artifacts in the processed speech and is usually more computationally expensive than simple frequency-shape processing. Instead, we propose to apply a simple time-invariant filter, which is a function of not only the power but also temporal aspects of the noise. More specifically, the proposed algorithm is based on replacing the long-term sub-band noise power $\sigma^2_{V_j}$ used in [7] with a long-term fractile noise power $\tilde{\sigma}^2_{V_j}$. The ϕ-fractile noise power $\tilde{\sigma}^2_{V_j}$ in the j-th band is here defined as

$$\text{Prob}\left\{|V(j,l)|^2 \leq \tilde{\sigma}^2_{V_j}\right\} = \phi, \qquad 0 \leq \phi \leq 1, \quad \forall l. \tag{7}$$

i.e., the threshold value $\tilde{\sigma}^2_{V_j}$ below which a fraction ϕ of the noise energy $|V(j,l)|^2$ in the j-th sub-band falls. This fractile noise power $\tilde{\sigma}^2_{V_j}$ then replaces the conventional noise power $\sigma^2_{V_j}$ in (1) to get a measure of the glimpses-aware fractile SNR $\tilde{\xi}_j$:

$$\tilde{\xi}_j = \frac{\sigma^2_{X_j}}{\tilde{\sigma}^2_{V_j}}, \qquad j = 1, ..., J. \tag{8}$$

The proposed algorithm replaces ξ_j in (4) with $\tilde{\xi}_j$ and $\sigma^2_{V_j}$ in (5) and (6) with $\tilde{\sigma}^2_{V_j}$.

Fig. 2 helps to understand how the long-term fractile noise power behaves differently in the presence of fluctuating and non-fluctuating maskers by showing the noise energy $|V(j,l)|^2$ in the j-th sub-band (j=15, whose central frequency is c_f=522 Hz) as a function of time l for both CS and SSN noise along with their corresponding long-term fractile noise power $\tilde{\sigma}^2_{V_j}$ and the conventional long-term noise power $\sigma^2_{V_j}$ as used by optimalASII. From the plot, we can observe that the fractile power of the CS is much lower than that of SSN, which is very close to the conventional noise power used in the original method. This, in turn, means that $\tilde{\xi}_j \simeq \xi_j$ in SSN, while $\tilde{\xi}_j \gg \xi_j$ in CS.

Fig. 3 shows the effect of the proposed use of fractile noise power by comparing the sub-band gains of optimalASII and the proposed method for both SSN and CS. We observe that in SSN (cf. Fig. 3(a, c)), the sub-band gains of optimalASII and the proposed method are very similar. On the other hand, in fluctuating CS (cf. Fig. 3(b, d)), the two methods use rather different gains, because the fractile SNR $\tilde{\xi}_j$ is much higher than the conventional SNR ξ_j and, consequently, no low frequency bands are severely suppressed.

4. RESULTS

We evaluate the performance of the proposed algorithm by comparison to relevant baselines. As baselines, we use the SEO algorithm [9], which was the best performing algorithm in HC-1 [2], and optimalASII [7], as the proposed algorithm is a modification of it. All speech intelligibility improvements provided by the methods are estimated using ESTOI [30] and STGI [31].

In the following, we assume that the proposed method decomposes speech and noise using a Hann window of length 512 samples with a 50% overlap, at a sampling frequency of 16 kHz. The number of sub-band used is set to $J = 64$ and the fractile $\phi = 0.15$.

4.1. Sensitivity to the fractile value ϕ

The proposed algorithm relies on the choice of a single parameter, namely the fractile value ϕ. Here, we demonstrate that the performance is insensitive to a wide range of choices for this parameter. To do so, the proposed algorithm is applied to the speech and noise data from HC-1 [2] for different fractile values ϕ and evaluated in terms of STGI and ESTOI, at SNRs between -30 and 10 dB.

Fig. 4 shows the performance of the proposed method as a function of the input SNR, for values of the fractile $\phi \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For the sake of space, only STGI scores are shown in the figure, as ESTOI predicts similar results. Here we observe that, as long as the fractile is chosen to be less than or equal to 0.5, the predicted intelligibility improvements provided by the proposed algorithm are very similar, thus suggesting an evident robustness to this parameter.

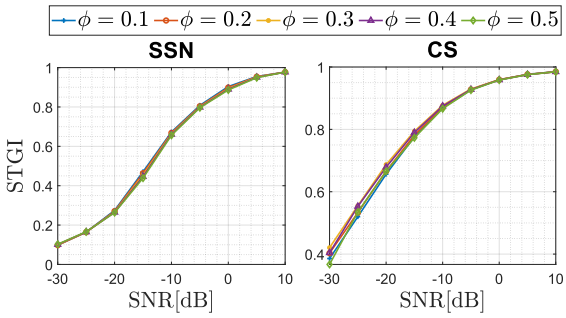

Fig. 4. Performance of the proposed method in terms of STGI for SSN (left column) and CS (right column) as a function of the SNR for different fractile values ϕ.

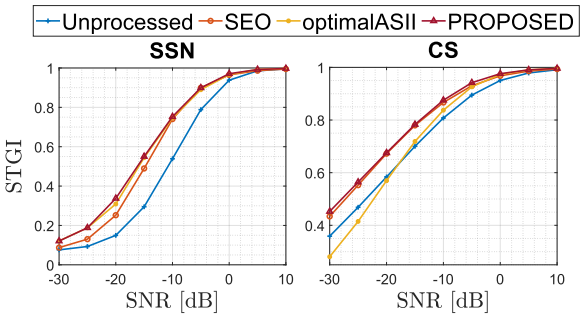

Fig. 5. Performance in terms of STGI as a function of the input SNR for speech and noise signals from HC-1. The left column shows the performance in SSN, the right column in CS.

4.2. Performance evaluation - Hurricane Challenge 1

In this experiment, we evaluate the performance of the proposed algorithm for speech and noise signals used in HC-1. Fig. 5 shows the results in terms of STGI for different input SNRs. We see that the proposed algorithm yields intelligibility improvements similar to optimalASII in SSN and similar to SEO, but much better than optimalASII, in CS. Secondly, in line with the listening experiments reported in the HC-1 [2], optimalASII and SEO exhibit similar performance in SSN, while SEO greatly outperforms optimalASII in CS. ESTOI scores are omitted for the same reasons as above.

4.3. Performance evaluation outside HC-1

The data used in HC-1 is limited, and only consists of a single male speaker uttering sentences from the Harvard corpus [32] and a single female speaker used to produce both SSN and CS.

The aim of this experiment is to evaluate the proposed algorithm for different speakers and noise conditions. To do so, we use a larger number of speakers drawn from three speech datasets, namely Dantale II [33], the American English Matrix Sentence test [34], and DARPA TIMIT [35]. As noise sources, we used SSN, whose long-term spectrum matches the average long-term spectrum of Dantale II, and competing speakers randomly drawn from the three speech datasets mentioned above.

Table 1 reports STGI scores for the three datasets, the two noise types, and 4 SNR conditions. The results show that the proposed algorithm can effectively deal with both non-fluctuating and fluctuating maskers, outperforming both baselines in every condition. ESTOI scores are omitted for the same reasons as above.

		SSN				CS			
Speech	Method	-15 dB	-10 dB	-5 dB	0 dB	-20 dB	-15 dB	-10 dB	-5 dB
Dantale	Unprocessed	0.153	0.316	0.603	0.846	0.457	0.555	0.679	0.771
	SEO [9]	0.508	0.725	0.857	0.925	0.618	0.710	0.808	0.868
	optimalASII [7]	0.677	0.822	0.902	0.953	0.574	0.687	0.802	0.876
	proposed	**0.684**	**0.825**	**0.908**	**0.956**	**0.649**	**0.735**	**0.833**	**0.891**
AEMST	Unprocessed	0.286	0.526	0.772	0.919	0.537	0.649	0.758	0.828
	SEO	0.546	0.733	0.854	0.921	0.616	0.719	0.810	0.865
	optimalASII	0.687	0.831	0.910	0.959	0.576	0.701	0.818	0.885
	proposed	**0.697**	**0.834**	**0.916**	**0.962**	**0.660**	**0.764**	**0.853**	**0.902**
TIMIT	Unprocessed	0.267	0.473	0.707	0.875	0.475	0.598	0.722	0.817
	SEO	0.516	0.726	0.857	0.918	0.591	0.702	0.803	0.869
	optimalASII	0.655	0.816	0.902	0.946	0.520	0.663	0.790	0.876
	proposed	**0.676**	**0.828**	**0.908**	**0.951**	**0.668**	**0.770**	**0.860**	**0.915**

Table 1. Performance in terms of STGI in SSN and CS at different SNRs for three speech datasets: Dantale II, American English Matrix Sentence test (AEMST), DARPA TIMIT.

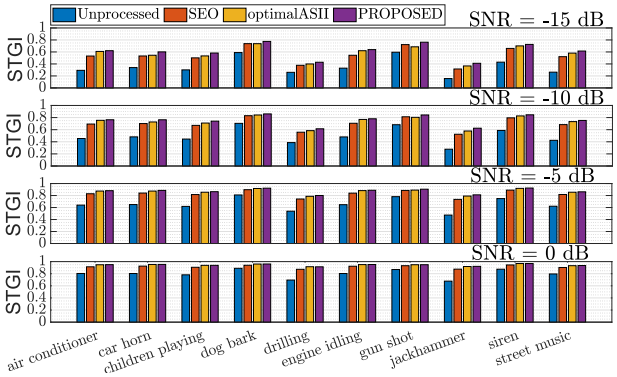

Fig. 6. STGI scores for clean speakers from Dantale II and the ten classes of noise in UrbanSound8K at four different SNRs.

In an additional experiment, we used environmental noise from the UrbanSound8K dataset [36]. The purpose of this experiment is to evaluate the algorithms when the long-term spectrum of the noise is not speech-like. The results of this experiment are reported in Fig. 6, which shows the STGI scores for each of the ten noise classes present in UrbanSound8K at four different SNRs. We see that the proposed method achieves higher STGI scores than the two baselines in every noise condition and SNR.

5. CONCLUSION

In this paper, we presented a novel near-end listening enhancement algorithm. The proposed algorithm is inspired by optimalASII [7] which aims at maximizing an approximation of the Speech Intelligibility Index under an energy constraint. However, unlike optimalASII, which relies on the long-term power of the background noise, we propose to use the long-term fractile noise power to accommodate the effect of "glimpsing" in fluctuating noises.

The proposed algorithm can discriminate between continuous and fluctuating noise maskers by using only long-term statistics of speech and noise. Despite its simplicity, the proposed method yielded speech intelligibility improvements, in terms of ESTOI and STGI, in every noise condition of the Hurricane Challenge 1 as well as for a wide range of speech and real-world environmental noise sources. Also, it performed similar to or better than relevant baseline algorithms for the tested acoustic scenarios.

Future studies will assess the proposed algorithm through subjective tests for speech intelligibility and listening effort.

6. REFERENCES

[1] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*. IEEE, 2006.

[2] M. Cooke et al., "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[3] M. Cooke et al., "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.

[4] M. Cooke et al., "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[5] É. Lombard, "Le signe de l'élévation de la voix," in *Annales des Maladies de l'Oreille, du Larynx du Nez et du Pharynx*, 1911, vol. 37, pp. 101–119.

[6] J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *J. Acoust. Soc. Am.*, vol. 29, no. 12, pp. 1320–1323, 1957.

[7] C.H. Taal et al., "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, 2013.

[8] E. Gentet et al., "Speech intelligibility enhancement by equalization for in-car applications," in *Proc. ICASSP*. IEEE, 2020.

[9] R. Takou et al., "Improvement of speech intelligibility by reallocation of spectral energy," in *Proc. Interspeech*, 2013, pp. 3605–3607.

[10] T.-C. Zorila et al., "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *13th Annual Conference of the ISCA*, 2012.

[11] R.C. Hendriks et al., "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Trans. Acoust., Speech, Sig. Proc.*, vol. 23, no. 5, pp. 851–862, 2015.

[12] E. Godoy and Y. Stylianou, "Increasing speech intelligibility via spectral shaping with frequency warping and dynamic range compression plus transient enhancement," in *Proc. Interspeech*, 2013, pp. 3572–3576.

[13] H. Brouckxon and W. Verhelst, "An overview of the VUB entry for the 2013 Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3602–3604.

[14] C. Chermaz and S. King, "A sound engineering approach to near end listening enhancement," in *Proc. Interspeech*, 2020, pp. 1356–1360.

[15] F. Bederna et al., "Adaptive compressive onset-enhancement for improved speech intelligibility in noise and reverberation," in *Proc. Interspeech*, 2020, pp. 1351–1355.

[16] V. Aubanel and M. Cooke, "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. Interspeech*, 2013, pp. 3592–3596.

[17] A.R. López et al., "Speaking style conversion from normal to Lombard speech using a glottal vocoder and Bayesian GMMs," in *Proc. Interspeech*, 2017, pp. 1363–1367.

[18] R. Biswas et al., "Optimal speech intelligibility improvement for varying car noise characteristics," *Journal of Signal Processing Systems*, vol. 94, no. 12, pp. 1429–1446, 2022.

[19] R. Biswas et al., "Statistically guided near-end speech intelligibility improvement through voice transformation and transfer learning," *IEEE/ACM Trans. Acoust., Speech, Sig. Proc.*, 2024.

[20] S. Seshadri et al., "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17230–17246, 2019.

[21] B. Bollepalli et al., "Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks," *Speech Communication*, vol. 110, pp. 64–75, 2019.

[22] H. Li et al., "iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning," in *Proc. Interspeech*, 2020, pp. 1336–1340.

[23] Y. Tang et al., "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Computer Speech & Language*, vol. 35, pp. 73–92, 2016.

[24] ANSI, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*, Acoustical Society of America, 1997.

[25] K.S. Rhebergen and N.K. Versfeld, "An SII-based approach to predict the speech intelligibility in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 115, pp. 2394–2394, 2004.

[26] E.C. Zsiga, *The Sounds of Language: An Introduction to Phonetics and Phonology*, John Wiley & Sons, 2012.

[27] C.A. Brown and S.P. Bacon, "Fundamental frequency and speech intelligibility in background noise," *Hearing Research*, vol. 266, no. 1-2, pp. 52–59, 2010.

[28] C. Spille and B.T. Meyer, "Listening in the dips: Comparing relevant features for speech recognition in humans and machines," in *Proc. Interspeech*, 2017, pp. 2968–2972.

[29] J.M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[30] J. Jensen and C.H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Acoust., Speech, Sig. Proc.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[31] A. Edraki et al., "A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction," in *Proc. Interspeech*, 2021, pp. 206–210.

[32] E.H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[33] K. Wagener et al., "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, 2003.

[34] B. Kollmeier et al., "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *Int. J. Audiol.*, vol. 54, no. sup2, pp. 3–16, 2015.

[35] J. Garofolo et al., "Darpa Timit acoustic-phonetic continuous speech corpus cd-rom {TIMIT}," 1993.

[36] J. Salamon et al., "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. on Multimedia*, 2014, pp. 1041–1044.